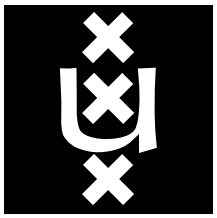


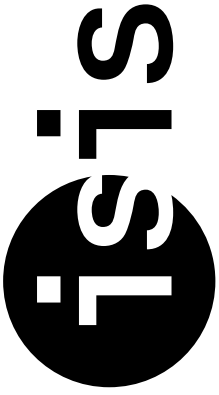
ISIS technical report series, Vol. 2001-13, October 2001

Searching in document images: what does the appearance of a document tell us about what it means?

Andrew Bagdanov, Marcel Worring and Arnold Smeulders
Intelligent Sensory Information Systems
Department of Computer Science
University of Amsterdam
The Netherlands

Abstract

The document understanding problem can be informally defined as the automatic extraction of meaning from documents. In the Intelligent Sensory Information Systems group we have experimented with analyzing the visual appearance of documents in order to extract meaning. That is, we concentrate on how documents look, rather than on what they say. We motivate this approach with several applications from document image understanding. First, we describe how document genre classification can be used to group visually similar documents together, which simplifies the analysis task for an entire class of documents. Second, we consider the logical block labeling problem. We show how logical labels (e.g. title, author, header, footer) can be assigned to blocks of text using a few visual features. Third, we discuss our approach to detecting the reading order of text using the visual structure of a document. The examples are based on the work in the field of content-based image retrieval (CBIR). Content-based image retrieval aims at searching and browsing image repositories on the basis of a visual specification of the query. The query may be one or preferably more examples, and the presentation may be a linear list of items, or prefarably a similarity grouping. Our research on colour representations of real world objects and colour composition of documents has shown that CBIR techniques can be successfully applied in order to simplify the document understanding problem.



Contents

Intelligent Sensory Information Systems
Department of Computer Science
University of Amsterdam
Kruislaan 403
1098 SJ Amsterdam
The Netherlands

tel: +31 20 525 7463
fax: +31 20 525 7490
http://www.science.uva.nl/research/isis

Corresponding author:
Andrew D. Bagdanov
tel: +31(20)525 7518
andrew@science.uva.nl
http://www.science.uva.nl/~andrew/isis

1 Introduction

> "*Before you become too entranced with gorgeous gadgets and mesmerizing video displays, let me remind you that information is not knowledge, knowledge is not wisdom and wisdom is not foresight. Each grows out of the other, and we need them all.*"
>
> Arthur C. Clarke (*Reinventing*, 1995)

There can hardly be any doubt that we are in the midst of an information revolution. The "information age" that pundits have long proclaimed is here. With it, we are forced to redefine our definition of, and interaction with, information and human knowledge. The advent of the personal computer, inexpensive mass storage devices, and the World Wide Web has shifted the emphasis from information access to *effective* access to information. There is a wealth of information available, but how can we find what we need?

Not only has the quantity of information exploded, but the variety as well. The traditional definition of "document" has grown to encompass a broad class of media including text, hypertext, video, color pictures, and sound. With this expansion come challenging new problems in information access. Traditional keyword and subject indexing fail to capture much of the information present in different types of media. Can traditional methods of indexing and retrieving information be modified to provide effective access to multimedia information?

Interestingly, even though digital information access has dramatically reshaped our understanding of information and information access, *paper* is still by and large the medium in which we conduct our public and personal lives. While many institutions have successfully reduced their reliance on paper–based communication, the "paperless office" is yet to be fully realized. Much of the information that moves through a typical office workflow is still paper–based. Still more information exists only in the paper archives of businesses and government institutions. Information technology has yielded powerful new modes of interaction with knowledge, but how can we provide effective access to the volumes of legacy information that do not exist in digital form?

These changes in the landscape of information and information access place library scientists and IT professionals in a unique position to shape the future of information access. In this article we address these questions of effective information access. We begin our discussion with an introduction to the concepts emerging from the field of content based information retrieval, which is concerned with providing effective access to pictorial archives. We then show how these ideas can be applied to the document understanding problem by characterizing documents according to how they look before trying to extract meaning. Finally, we look at several concrete applications of these ideas.

2 Content Based Image Retrieval

Content based image retrieval is a relatively young field of study, barely ten years old [4]. It emerged from three formerly unrelated disciplines: computer vision, information retrieval, and databases. Its goal is to provide effective access to large image databases from many different domains. Figure 1 illustrates a simple query model for an image retrieval system.

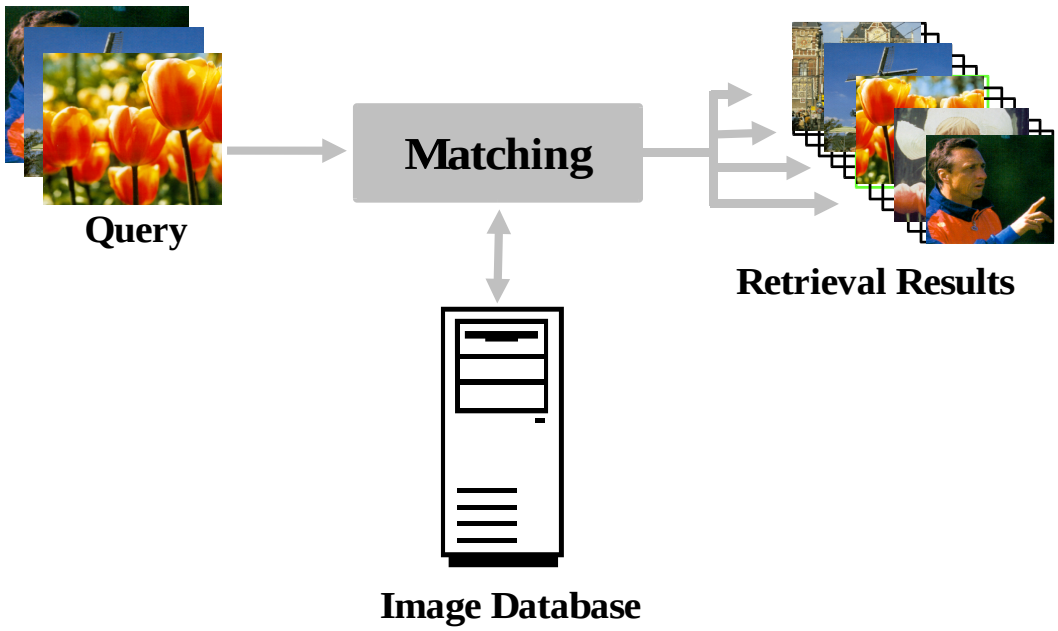


Figure 1: The content based image retrieval query model. An example image is provided to the retrieval system, which matches it against a large database of archived images. A (ranked) list of similar images is returned.

In this section we give a brief introduction to content based image retrieval, and discuss the aspects that set it apart from other types of information retrieval.

2.1 Image Retrieval

Consider the image shown in figure 2. Take a few minutes to try to describe the image. Some possible descriptions might be:

- It is a picture of a desert landscape.
- It is an image of a desert landscape showing sand dunes.
- It is a scene picturing a typical Death Valley landscape.
- It is a picture of a Creosote Bush.
- It is picture of *Larrea tridentata.*
- It is a scene showing a typical Death Valley landscape, which is representative of desert landscapes in the southwestern portion of the United States. Shown are sand dunes, a Creosote Bush (*Larrea tridentata*), and the clear blue sky of an autumn day.

The point to this familiar exercise is that there is something about an image that cannot be captured with words. It must be seen to communicate all of its qualities. The retrieval of visual information differs from other information retrieval problems in that visual response

Figure 2: A desert landscape of some sort.

is innately visceral. It is difficult, if not impossible, to quantify all of the interesting aspects of an image.

Images, unlike other types of information, are not always produced under rigidly controlled circumstances. Imaging conditions may create unwanted artifacts in the image, often referred to a *noise* in the image processing community, that are not related to the content of the image. In figure 3 the same box of boerencake mix is imaged under different conditions. This figure illustrates some important factors that affect the digital representation of an imaged scene. For any scene there can be literally millions of different data arrays that can be imaged. A direct consequence of this is the futility of comparing the raw data of images to determine content similarity.

This example motivates the need for *invariance* in content based image retrieval [3]:

Definition 1 *A feature F is* **invariant** *to condition A if, for object x, F produces the value $F(x)$ regardless of the effect of condition A.*

The concept of invariance expresses our desire that an algorithm performed on an image should return identical results under varying imaging conditions. Several of the images in figure 3 were acquired under different illumination conditions. In this case we would desire features that are invariant to differences in illumination. Some of the boxes in figure 3 are also occluded, and we would also want features that are invariant to some degree to occlusions. There is an important tradeoff between invariance and discriminative ability. We could design an "algorithm" that classified every image as *Boerencake Mix*, which would be completely invariant to all imaging conditions but lack any discriminative power for differentiating, for example, between images of Boerencake Mix and snapshots of my cat Mojo.

Figure 3: Some of the factors affecting image acquisition. The object in each photo is the same, but the digital representation varies drastically depending on illumination, occlusion, orientation, etc.

2.2 The Semantic Gap

When we look at the images in figure 3 we can immediately tell that the object in each image is identical. The human content recognition "algorithm" is remarkably robust, which brings us to what is in a sense the crux of the content based image retrieval problem:

Definition 2 *The* **semantic gap** *is the lack of coincidence between the information that one can extract from the (visual) data and the interpretation that the same data has for a user in a given situation.*

The semantic gap expresses the inherent difference between the digital representation of an image and the interpretation that a user associates with it. A user may be interested in retrieving interesting pictures of desert landscapes, but the computer only "knows" that it has pictures with some brown in them. Much of the research in content based image retrieval concentrates on bridging the semantic gap using invariance – particularly color invariance [3], and user interaction [6].

3 CBIR and Document Understanding

In this section we describe how concepts from content based image retrieval may be applied to the document understanding problem. Written communication is highly structured, and we can leverage this fact in applying techniques from content based image retrieval.

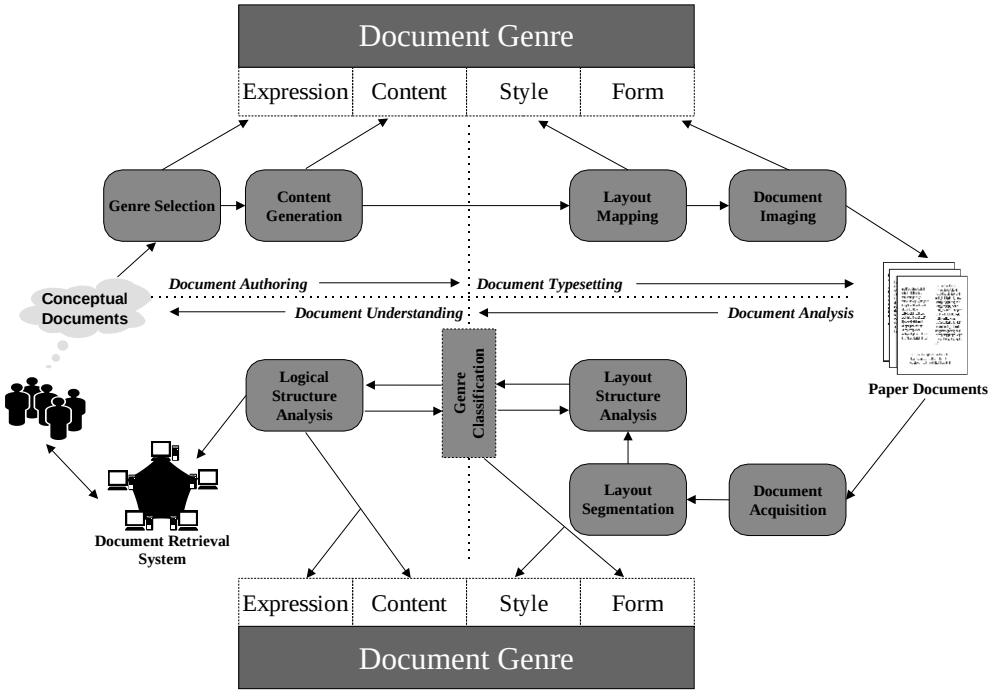


Figure 4: The modern document lifecycle. A document evolves through the authoring and typesetting phases. The document analysis and understanding processes incrementally reconstruct the meaning of the document. The end result is then indexed into a document retrieval system.

Images are generally unconstrained and highly unstructured, while machine printed documents utilize rhetorical, logical, and physical structure in order to effectively communicate a message. We look first at the general problem of document image understanding.

3.1 Document Understanding

The task of a complete document analysis system, as shown in figure 4, is to take a binarized scan of a paper document and accurately reconstruct the logical meaning of the document for indexing in a document retrieval system. This process can be thought of as the inverse of document authoring, and is typically divided into two stages: document image analysis, and document image understanding. Document image analysis includes scanning up to layout component identification. Document understanding encompasses all of the procedures used to transform layout components detected in the document analysis stage into an accurate representation of the content of the document.

In most forms of written communication there are specific logical information elements that are expected to be present. For example, a business letter is expected to have a date, from–address, to–address, opening, body, and closing. These logical elements are transformed in the typesetting phase into physical layout structures such as section headings, titles, and bibliographic references. Layout structures *usually* conform to accepted pro-

tocols and standards that allow a reader to accurately reconstruct the intended message (publications such as *Wired* magazine being notable exceptions).

3.2 The Semantic Gap and Document Understanding

The highly structured nature of machine printed documents aside, there is still a huge gap between the raw array of pixels acquired by a document scanner and the conceptual document the author had in mind when originally creating it. Even accurate optical character recognition is not sufficient in bridging the semantic gap in document understanding. It is necessary to associate *meaning* to the identified elements of a document image. This is the primary focus of document image understanding, and below we provide several examples of how visual structure can be used to effectively bridge the semantic gap.

4 Case Studies

In this section we discuss several applications of the ideas presented above. Out research is primarily focused on bridging the semantic gap in document understanding by analyzing the visual structure of documents.

4.1 Document Genre Classification

Document understanding is difficult due to the wide diversity of possible document types, or *genres*, within the domain of document processing systems. Even in the restricted domain of machine printed documents, there exists a daunting variety of document types.

Definition 3 *A* **document genre** *is a category of documents characterized by similarity of expression, style, form, or content.*

Figure 4 illustrates how document genre incrementally evolves through the document creation process, and is then reconstructed through document understanding. We are primarily interested in the the *form* and *style* components of document genre, as these represent the visual aspects of document layout. More precisely, we are interested in determining how the visual components affect the interpretation of a document. Figure 5 illustrates the document genre classification process.

Our research has shown that many documents types can be distinguished from visual appearance alone [2]. Applications of such genre classification techniques include automatic routing of documents in office workflows, and the improvement of logical information extraction for specific document genres. Experiments have shown that document genre can be correctly identified about 97% of the time.

4.2 Segmentation of Complex Color Layouts

Focus has been changing in the document analysis research community toward processing of more and more complex document layouts. Until now most document analysis has centered on relatively simple black and white documents with simple layout styles. Figure 6 illustrates the changing complexity of document layout considerations.

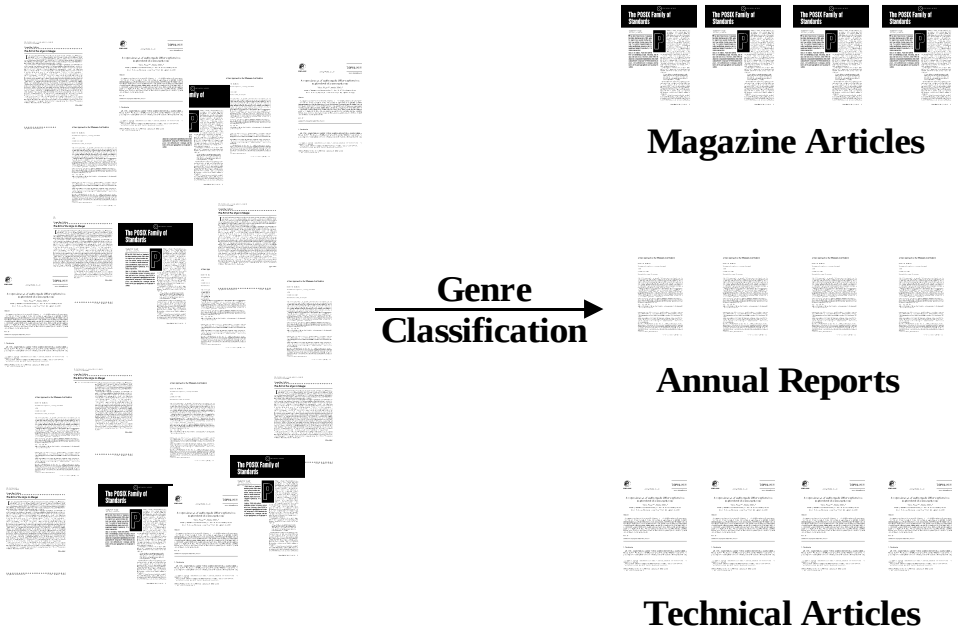


Figure 5: Genre classification based on visual similarity. On the left is an unsorted pile of documents which are then grouped according to visual similarity. Many classes of documents are distinguished by their visual appearance. Genre classification during the early stage of document image analysis can help to identify the types of logical information present, and narrow the range of processing steps needed at higher levels.

Our research has shown that complex color documents can be effectively segmented by analyzing the three dimensional color structure of scanned document images [7].

4.3 Reading Order Detection

Accurate determination of the reading order of detected text regions is an essential component of document understanding. The problem becomes particularly acute when processing complex layouts such as magazines and newspapers. Figure 7 illustrates the ambiguity of reading order. Once text regions have been identified, there may be many possible reading orders for the collection of detected rectangles.

Fortunately there are many accepted conventions for reading order, and in many cases the correct order may be determined by looking at the relative positions and sizes of the detected text blocks. Some ambiguities cannot be resolved using visual structure alone, however, and it is necessary to employ natural language processing to determine the correct reading order [1].

4.4 Logical Block Labeling

Figure 8 illustrates the interaction of all of the document understanding components we have discussed. The last task in the process is the assignment of logical labels (e.g. title, caption, page number) to each of the detected text regions. Many approaches to this problem rely on accurate OCR of the text in each region in order to assign a logical label.

Figure 6: The increasing demands on document understanding. As the complexity of modern document layout continues to increase, so do the demands on document understanding systems. In the past ten years the focus of much document analysis research has shifted toward processing more and more complex styles of documents.

We have experimented with using a small number of visual features to discriminate between logical classes [5]. Logical elements such as title, caption, and page number can be recognized with an accuracy of approximately 91%.

5 Conclusions

Our research has shown that concepts and approaches from content based image retrieval can augment existing approaches to document understanding by bridging the semantic gap between conceptual document and a scanned document image. By pre–classifying documents into genres according to visual appearance it is possible to specialize document understanding approaches to particular document classes. Through the analysis of the color composition of complex document layouts we have shown that text can be correctly identified and segmented in such documents. By combining spatial and linguistic information we have been able to accurately determining the reading order of text in complex documents. Lastly, by combining the above approaches we have shown that logical meaning can be extracted from documents by using visual characteristics.

While our research has been concentrated on machine printed documents, it should be noted that the techniques described above can be generalized to electronic documents (i.e. HTML, PDF, Word, etc.) as well. This is likely to be an active area of research in the near future, as more and more documents are being created and published in electronic form

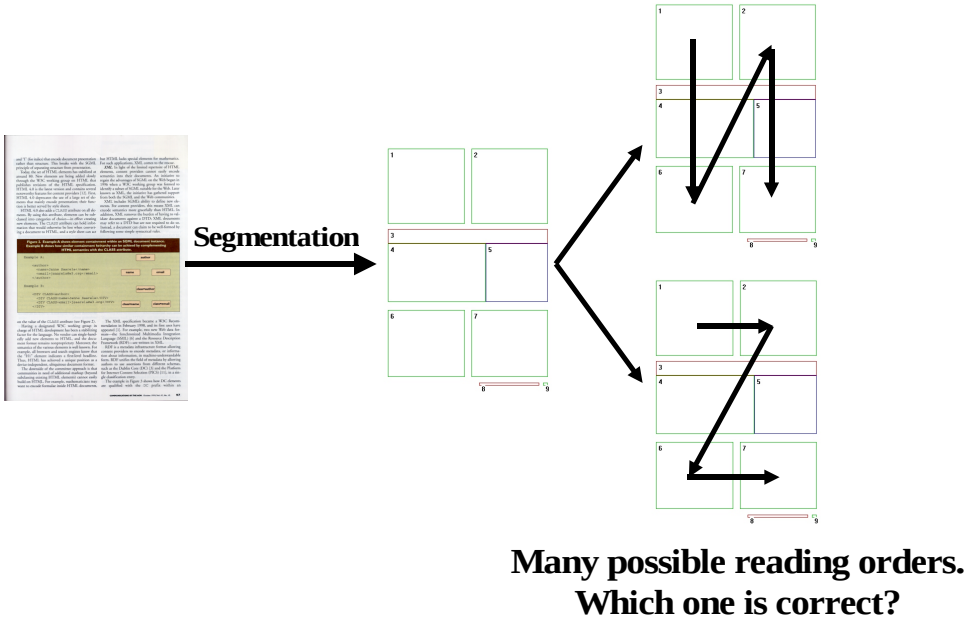


Figure 7: The ambiguity of reading order. For a single document segmentation there may be many possible reading orders. The segmentation of a document only increases the level of abstraction. Instead of reasoning about sequence of unstructured bits, we can now reason about a collection of semi–structured rectangles. The semantic gap is narrowed, but not completely bridged.

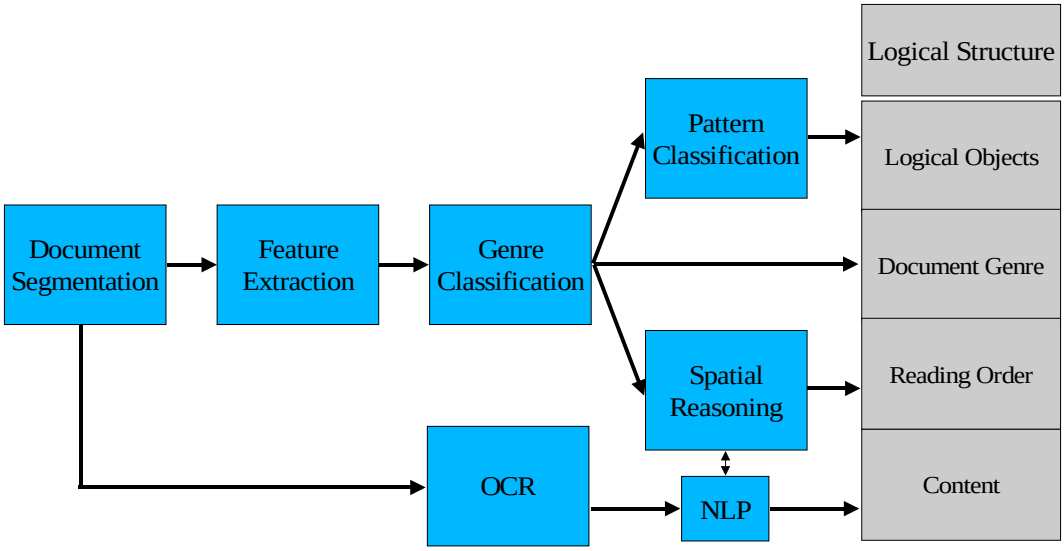


Figure 8: The complete document understanding process. By combining visual, typographical, and content–based features the logical document structure is reconstructed. The goal is to approximate the conceptual document that the author had in mind.

exclusively.

References

[1] M. Aiello, C. Monz, and L. Todoran. Combining linguistic and spatial information for document analysis. In *Proceedings of the RIAO'2000 Content-Based Multimedia Information Access*, pages 266–275, April 2000.

[2] A. Bagdanov and M. Worring. Fine-grained document genre classification using first order random graphs. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR2001)*, pages 79–83, September 2001.

[3] T. Gevers and A. Smeulders. Pictoseek: Combining color and shape invariant features for image retrieval. *IEEE transactions on Image Processing*, 9(1):102–119, January 2001.

[4] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 2000.

[5] L. Todoran, M. Aiello, C. Monz, and M. Worring. Logical structure detection for heterogeneous document classes. In *SPIE Proceedings Series: Document Recognition and Retrieval VIII*, pages 99–110, January 2001.

[6] M. Worring, A. Smeulders, and S. Santini. Interaction in content-based retrieval: an evaluation of the state-of-the-art. In R. Laurini, editor, *Advances in Visual Information Systems*, number 1929 in LNCS. 2000.

[7] M. Worring and L. Todoran. Segmentation of color documents by line oriented clustering using spatial information. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition (ICDAR'99)*, pages 67–70, September 1999.

ISIS reports

This report is in the series of ISIS technical reports. The series editor is Rein van den Boomgaard (reports-isis@science.uva.nl). Within this series the following titles are available:

References

[1] Theo Gevers. Robust photometric invariant region detection in multispectral images. Technical Report 2001-06, Intelligent Sensory Information Systems Group, University of Amsterdam, May 2001.

[2] Jan-Mark Geusebroek, Frank J. Seinstra, and Arnold W.M. Smeulders. Exhaustive orientation scale-space computation - a robust approach to curvilinear structure detection. Technical Report 2001-07, Intelligent Sensory Information Systems Group, University of Amsterdam, May 2001.

[3] Marcel Worring and Theo Gevers. Interactive retrieval of color images. Technical Report 2001-08, Intelligent Sensory Information Systems Group, University of Amsterdam, June 2001.

[4] Jan-Mark Geusebroek, Minh Anh Hoang, Jan van Gemert, and Marcel Worring. Searching for images in biomedical publications. Technical Report 2001-09, Intelligent Sensory Information Systems Group, University of Amsterdam, July 2001.

[5] Andrew D. Bagdanov and Marcel Worring. First order gaussian graphs for efficient structure classification. Technical Report 2001-10, Intelligent Sensory Information Systems Group, University of Amsterdam, August 2001.

[6] Sennay Ghebreab and Arnold W.M. Smeulders. Hypersnakes: Active stochastic object models. Technical Report 2001-11, Intelligent Sensory Information Systems Group, University of Amsterdam, September 2001.

[7] Sennay Ghebreab and Arnold W.M. Smeulders. Medical image segmentation by hypersnakes. Technical Report 2001-12, Intelligent Sensory Information Systems Group, University of Amsterdam, September 2001.

[8] Andrew D. Bagdanov, Marcel Worring, and Arnold Smeulders. Searching in document images: what does the appearance of a document tell us about what it means? Technical Report 2001-13, Intelligent Sensory Information Systems Group, University of Amsterdam, October 2001.

You may order copies of the ISIS technical reports from the corresponding author or the series editor. Most of the reports can also be found on the web pages of the ISIS group (http://www.science.uva.nl/research/isis).

Intelligent Sensory Information Systems
University of Amsterdam
The Netherlands